Service and Data Challenges Schedule and Status

Ian Fisk

CMS Coordination Meeting

April 11, 2006

Period of Transition

CMS has a lot a new components to work with this year

- New Grid stacks from LCG (gLite 3.0) and OSG (0.4.1)
- New Framework for CMS (CMSSW)
- New Data Management Components (DBS and DLS)
- New Production Tools for simulation submission (ProdAgent)
- Continuing development on CRAB (1.1)

Most of these are redesigns with experience from the previous generation

- Need to work our way back to the same level of functionality in a few cases

There is a lot of work to arrive at a fully functional system by the end of the year.

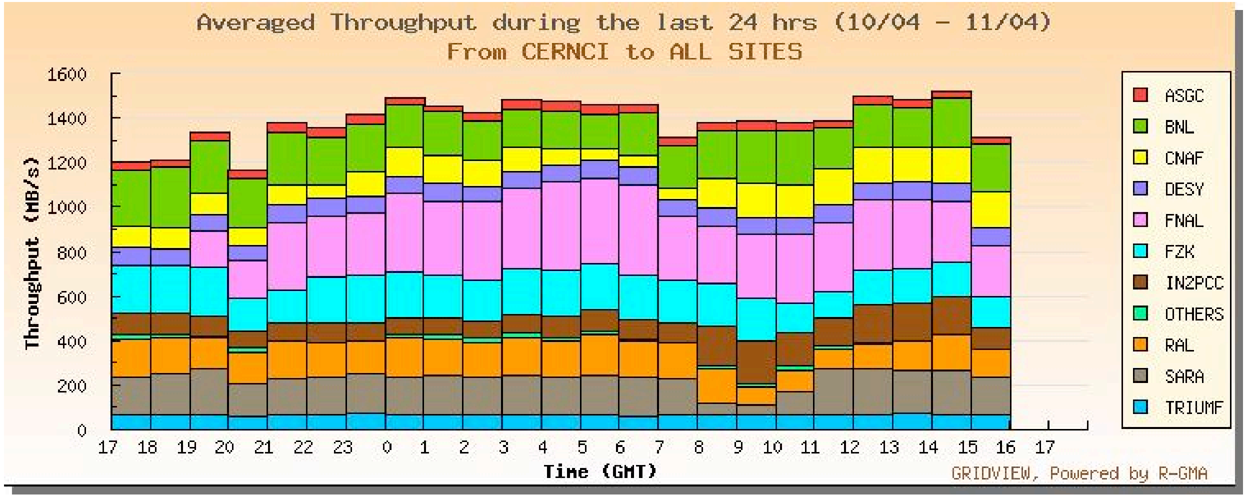
TimeLine (1/5)

April 5-20 SC4 Throughput Phase

- Initial Disk-to-Disk Transfers
- FNAL Expectations 200MB/s sustained
    - 10% degradation over 24 hours requires an explanation
- After Easter Tape Transfers begin
- Initial request of FNAL was 75MB/s sustained
    - We requested to use the CMS planning numbers that call for 50MB/s
    - Not clear if the new staging pools may have cleared much of the issues
    - Should discuss an improved priority on the shared STKEN resources for a couple of days
- Tier-1 to Tier-2 transfers follow
- In US-CMS we would like to demonstrate disk-to-disk transfers at 50% of the available network
    - Measure base and then calculate available network
    - So far even 2Gb/s (250MB/s) is proving very difficult
        - Local dCache load and Tier-2 configuration

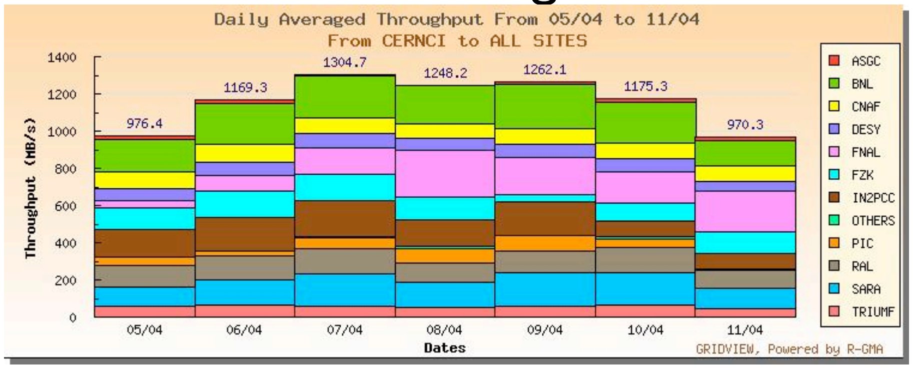
Status of Throughput Challenge

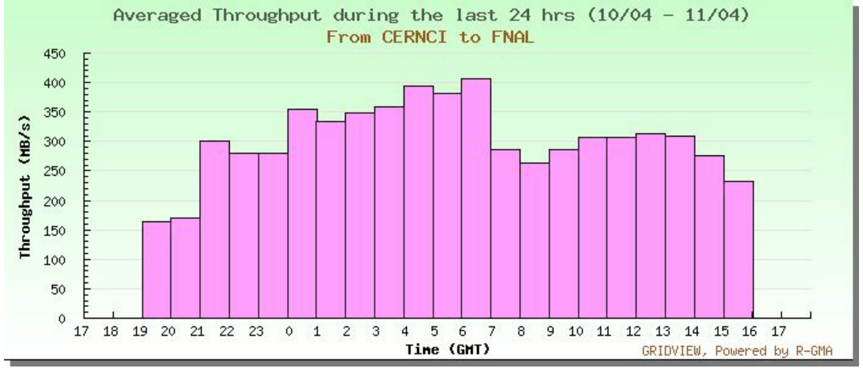
Transfers last 24 hours



Takes two ends for a successful transfer

- ➨ Early problems were ours, later were theirs
  - Downtime, configuration issues, and cost model, later FTS srmCopy

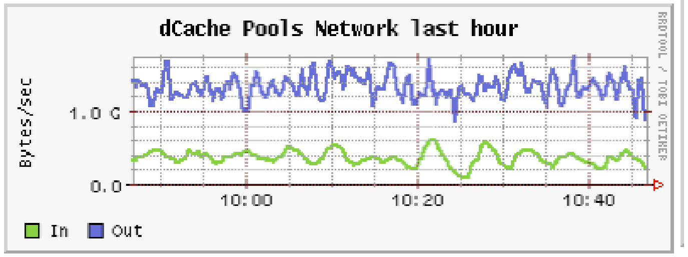




Throughput Status Cont.

The transfer challenge continues to be effort intensive on both sides of the transfer

- Tools to drive transfers fail or hang. Corrected with a series of cron jobs and scripts
  - This is the first large scale test of srmCopy transfers
    - Resulted in big improvements to FTS
- Still big fluctuations in the number of transfers and the resulting rate



On the FNAL side

- Monitoring is lacking and tracking problems is time consuming and requires expertise
  - System can fail in a lot of places, so quick triage is important
  - Discovering black holes before they kill transfers
  - Load balancing - cost model for destination sites

Next Steps for Throughput

We'll have to recycle a lot of tape, but we have succeeded at the tape transfers in the past

- ➡ Rate has increased marginally.

The Tier-1 to Tier-2 transfers are more likely to require effort

- ➡ Outside any discussion of data management or data access, the underlying SRM transfers from dCache to dCache has to be able to drive network high performance network links
  - ● Essentially all the Tier-2s will be connected at 10Gb by the end of the calendar year
    - In general access will be bursty
  - ● The limiting factors are the end points. Need to work on configuration and architecture with the Tier-2 centers
    - It is the same exercise we perform with the Tier-0 to Tier-1 challenge and should be easier with each iteration
    - Nonetheless hard work
    - Require central support from FNAL for the Tier-2

Timeline (2/5)

May 1 - June 1 Grid Roll-out

- OSG will roll out 0.4.1 ~May 1
  - Includes support for web services, but probably not by default
    - Incremental changes
- gLite 3.0 on the LCG side
  - More disruptive
  - Includes the gLite RB based on Condor-C
    - Still in basic functionality testing
    - Release candidate 2 was made today
  - Schedule calls for shake out tests for next few days
  - Hopefully scale tests by the end of April
  - May - June roll out onto production resources

Grid Deployment

OSG 0.4.1 will be deployed on the ITB and later production resources

- ➨ Process is well established. 0.4.0 upgrade was fairly smooth
- ➨ We see scaling issues at several of the Tier-2 centers and FNAL regarding the job execution environment
  - LCG invested more effort at eliminating shared file system
  - OSG is catching up, but we need to propagate the deployment
    - New development of managed fork queue is good
  - Push for more use of Storage Elements

gLite 3.0 is currently deployed in a PreProduction Mode (PPS)

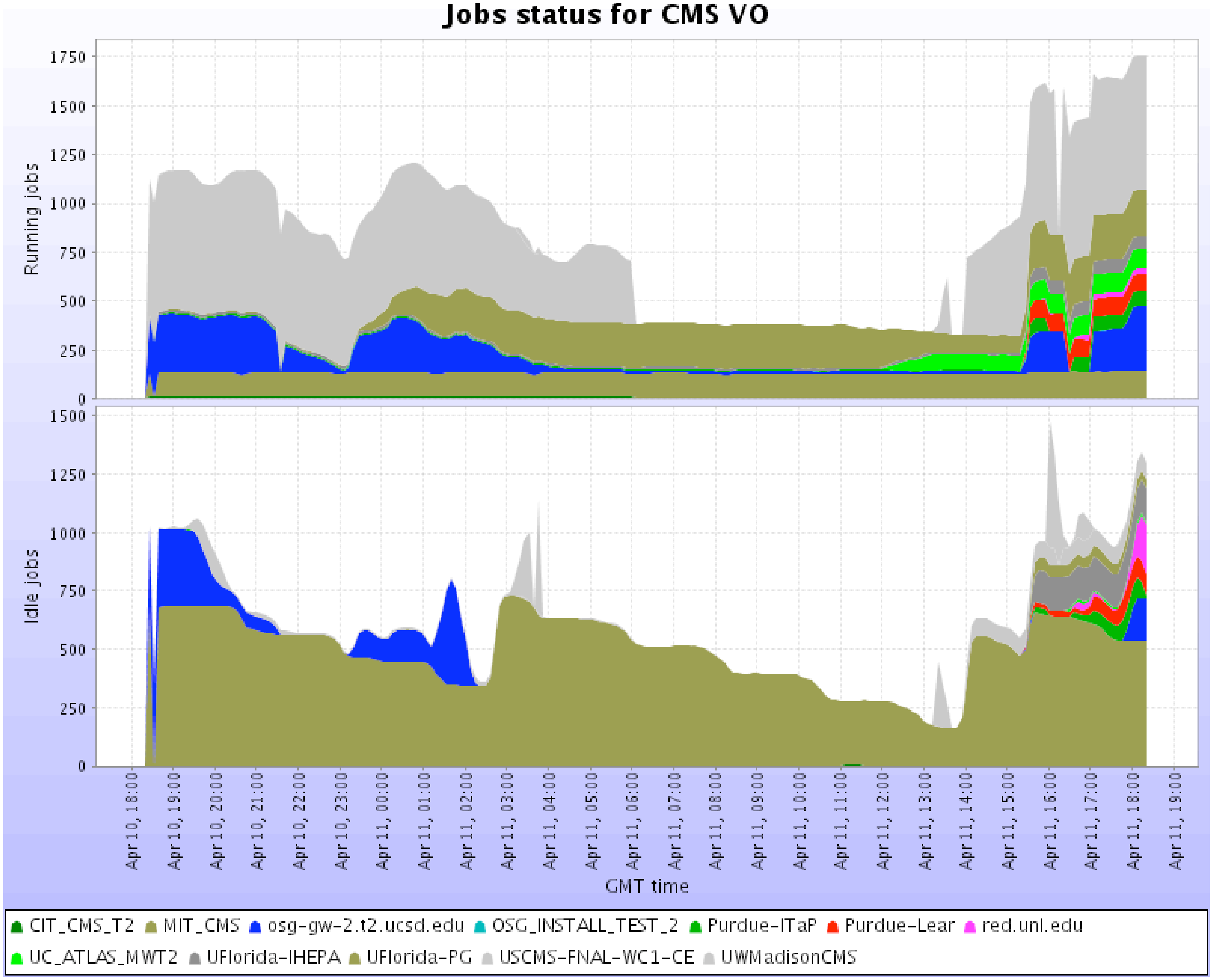
- ➨ Need to get PPS setup at FNAL
  - For CMS to exercise gLite in realistic ways we need to access sufficient processing resources and quantities of data
    - Submit through PPS service to production batch slots
    - A lot of CMS data access is through local protocols. Need to submit processing requests to production storage

Testing

CMS is developing a group of robots and end-to-end monitoring systems

- Robots handle bulk submission to resources.
  - Failures are cataloged and a sudden increase indicates a problem
  - Also generates the load for stress testing
    - Currently used for LCG 2.7 and OSG 0.4.0
      - Used to get failure statistics on both grids
      - Used to overload the UNL NFS server
- A heart beat monitor was introduced for transfers
  - http://cmsdoc.cern.ch/cms/aprom/TransferHeartbeat/browser
  - Generates transfers between sites using SRMCP
    - Keeps logging information
    - Maintains site status
  - Later in the month we will progress to some low level transfer load tests
    - Goal is 20MB/s between sites

Grid Plots


Jobs status for CMS VO
Running jobs
1750
1500
1250
1000
750
500
250
0
Idle jobs
1500
1250
1000
750
500
250
0
Apr 10, 18:00
Apr 10, 19:00
Apr 10, 20:00
Apr 10, 21:00
Apr 10, 22:00
Apr 10, 23:00
Apr 11, 00:00
Apr 11, 01:00
Apr 11, 02:00
Apr 11, 03:00
Apr 11, 04:00
Apr 11, 05:00
Apr 11, 06:00
Apr 11, 07:00
Apr 11, 08:00
Apr 11, 09:00
Apr 11, 10:00
Apr 11, 11:00
Apr 11, 12:00
Apr 11, 13:00
Apr 11, 14:00
Apr 11, 15:00
Apr 11, 16:00
Apr 11, 17:00
Apr 11, 18:00
Apr 11, 19:00
GMT time
CIT_CMS_T2
MIT_CMS
osg-gw-2.t2.ucsd.edu
OSG_INSTALL_TEST_2
Purdue-ITaP
Purdue-Lear
red.unl.edu
UC_ATLAS_MWT2
UFlorida-IHEPA
UFlorida-PG
USCMS-FNAL-WC1-CE
UWMadisonCMS

Timeline (3/5)

June 1-14: Service Challenge Running

- In SC3 we had a number of goals designed to exercise various aspects of the computing model
  - In part due to unstable or unavailable external and CMS components we concentrated on
    - Transferring data to Tier-1s and TIer-2s, publishing its existence, and accessing with analysis applications.
- In SC4 we would like to repeat to demonstrate a fuller set of computing model activities
  - Dress rehearsal for CSA06
  - We need to hit 25% of the system scale in the beginning of the year and closer to 50% by the end

Operational Goals 2008

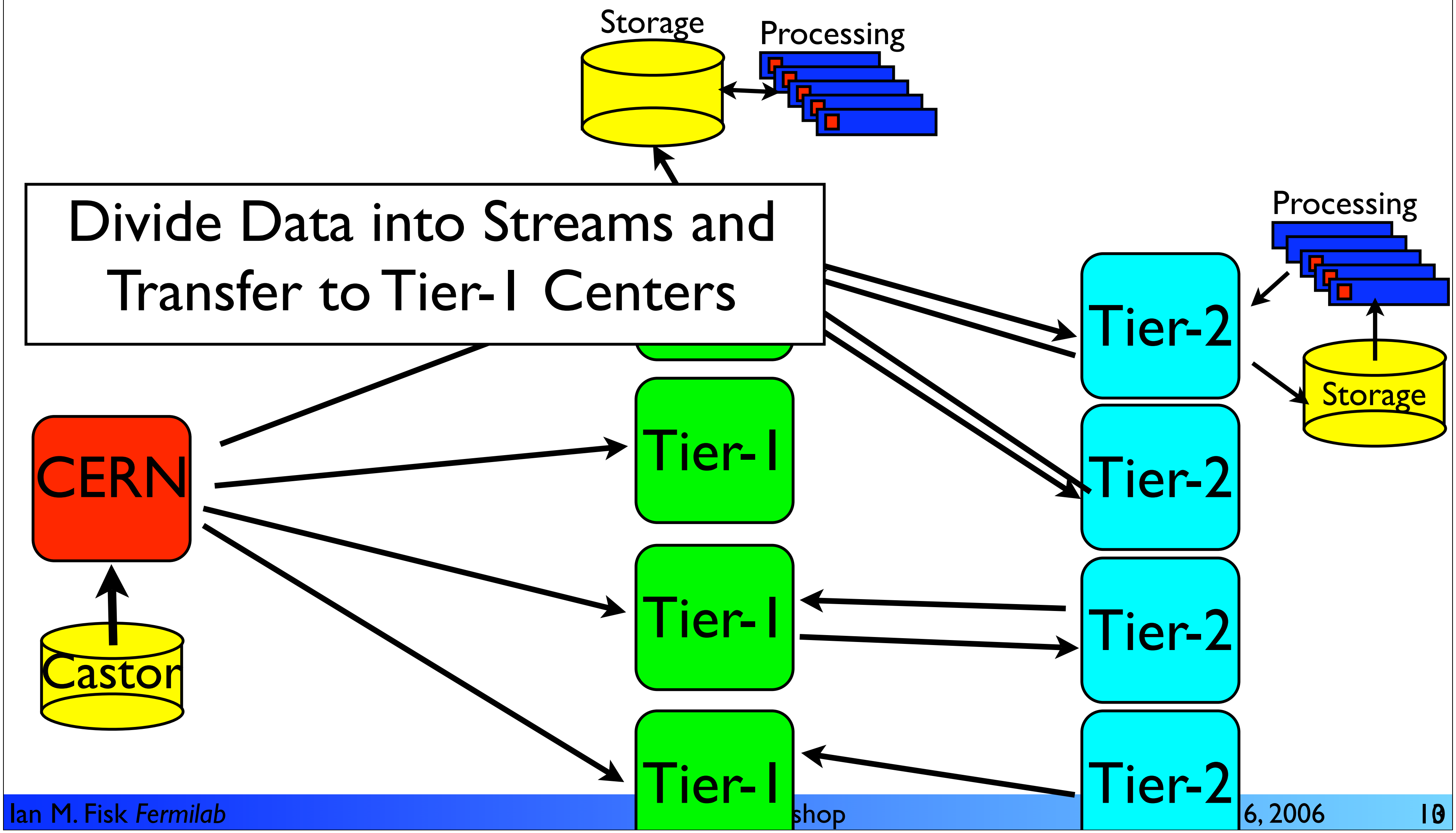
CMS needs to be at production scale services in 2008

- ➨ Network transfers between T0-T1 centers
  - ● 2008 scale is roughly 600MB/s
- ➨ Network transfers between T1-T2 centers
  - ● 2008 Peak rates from Tier-1 to Tier-2 of 50-500MB/s
- ➨ Selection Submissions and Transfers to Tier-1 centers
  - ● 2008 submission rate 50k jobs per day to integrated Tier-1 centers
- ➨ Analysis Submissions to Tier-2 Centers centers
  - ● 2008 Submission rate 150k jobs to integrated Tier-2 centers
- ➨ MC Production jobs at Tier-2 centers
  - ● 2008 rate is 1.3 x10^9 Events per year

Assuming we cannot easily more than double the scale each year, we should be able to demonstrate 25% of the 2008 scale now

- ➨ Divide the above by 4 now (2 by the end of the year)

Tier-0 to Tier-1 Basic Flows



Divide Data into Streams and Transfer to Tier-1 Centers

Tier-1 Transfers

In 2008 CMS expects ~300MB/s being transferred from CERN to Tier-1 centers on average

- Provision roughly twice that
- Assume peaks to recover from downtime and problems with a factor of two

Demonstrate aggregate transfer rate to 300MB/s sustained on experiment data by the end of year to tape at Tier-1 by the end of the challenge

- 150MB/s in the spring

At the start of the experiment, individual streams will be sent to each Tier-1 center

- During the transfer tests, data to each tier-1 will likely have a lot of overlap

Pieces Needed for CMS

Phedex integration with FTS (File Transfer Service) (INFN Contribution)

- ➨ We have a prototype, though not scale validation

CMS Data Management Prototype (FNAL Participating)

- ➨ We have a basically functional DBS (dataset bookkeeping service) and DLS (dataset location service) (query capabilities and discovery are lacking)

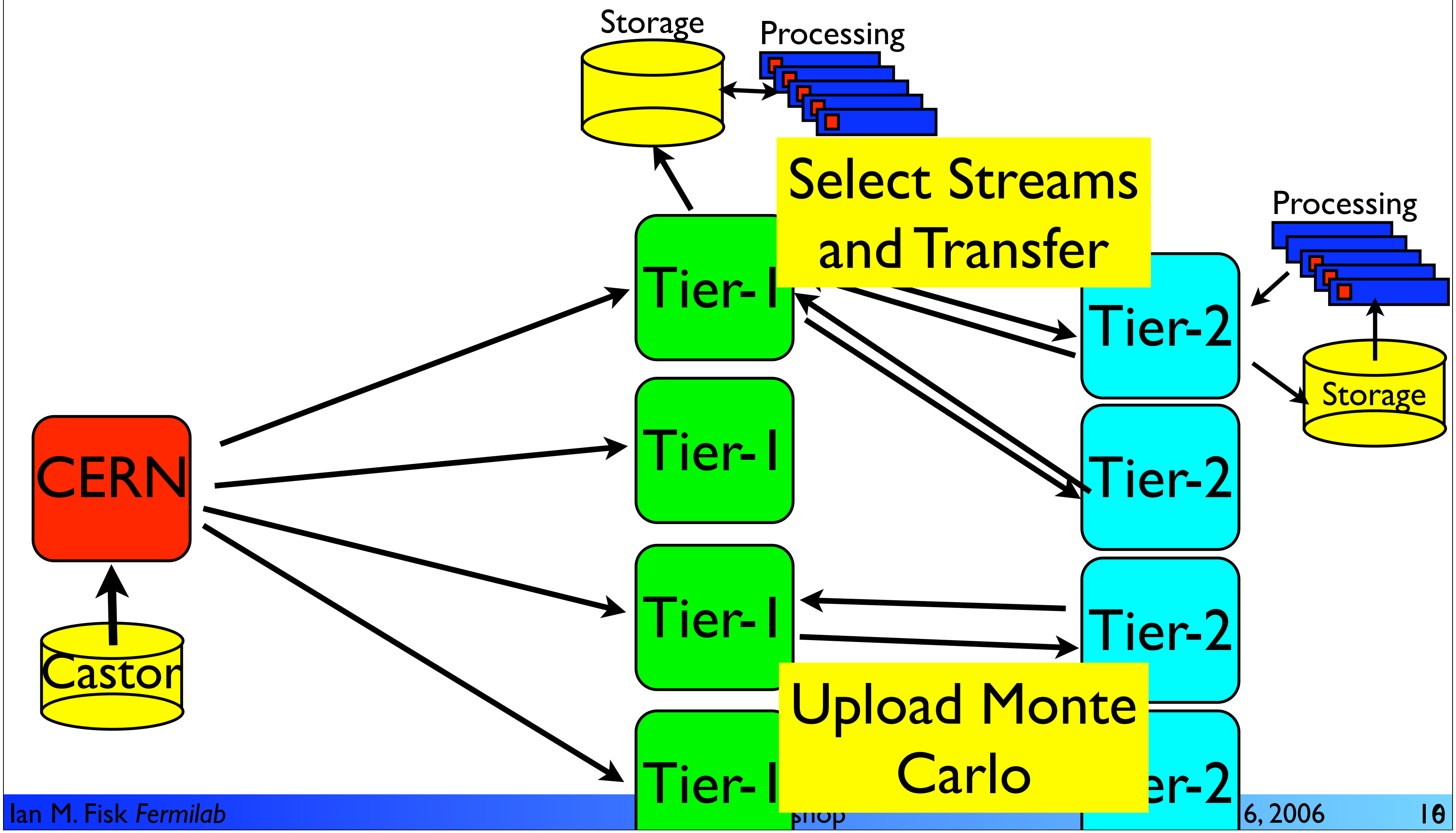
New Event Data Model Data (FNAL Participating)

- ➨ Expect our first 10TB sample roughly on April 1 (1TB now)
- ➨ The ability to transfer data on files is a component, but the experiment needs to transfer defined groups of files, validate integrity and make them accessible.

SC3 Rerun clearly demonstrated capabilities of transfer at this rate to disk

- ➨ CMS would like to replicate multiple copies of 10TB sample from Tier-0 to Tier-1 tape at a rate of 150MB/s
  - We would also like to exercise getting them back for applications

Tier-1 to Tier-2 Basic Flows



Select Streams and Transfer

Upload Monte Carlo

CMS Pieces

CMS Tier-1 to Tier-2 transfers in the computing model are likely to be very bursty and driven by analysis demands

- ➡ Network to Tier-2 are expected to be between 1Gb/s to 10Gb/s assuming 50% provisioning and a 25% scale this spring
  - ● Desire is to reach ~10MB/s for worst connected Tier-2s to 100MB/s to best connected Tier-2s in the Spring of 2006.

The Tier-2 to Tier-1 transfers are almost entirely fairly continuous simulation transfers

- ➡ The aggregate input rate into Tier-1 centers is comparable to the rate from the Tier-0.
  - ● Goal should be to demonstrate 10MB/s from Tier-2s to Tier-1 centers
    - 1TB per day
    - The production infrastructure will be incapable of driving the system at this scale, so some pre-created data will need to be used.
  - ● FNAL contributing to Production Infrastructure

US Tier-2s are in good shape and will be on the front line

Schedule and Rates

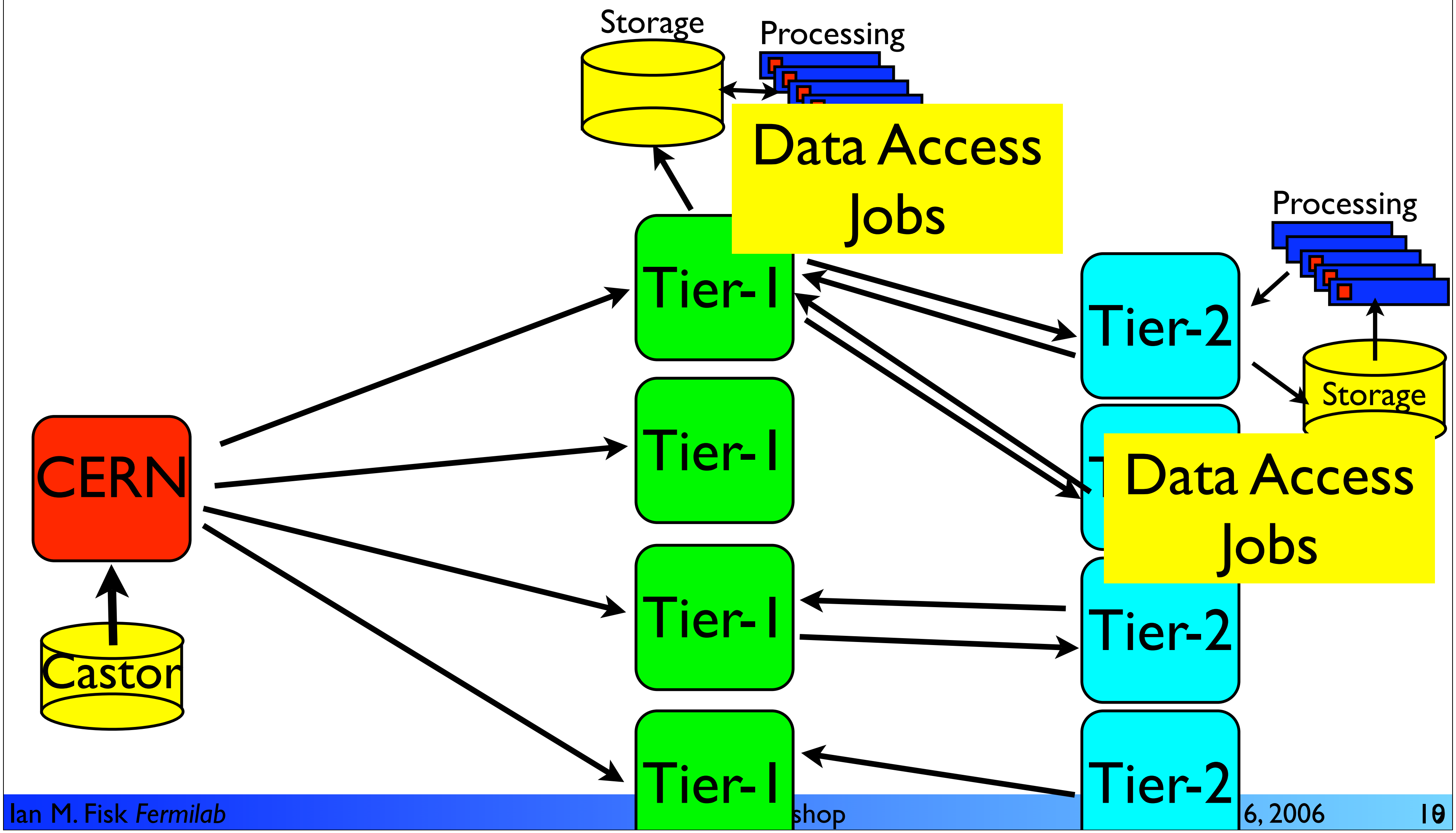
Schedule Items

- March 15 FTS driven transfers from PhEDEx
- Starting in April CMS would like to drive continuous low level transfers between sites that support CMS
  - On average 20MB/s (2TB per day)
  - The PhEDEx heartbeat system is functional all sites should register
    - http://cmsdoc.cern.ch/cms/aprom/TransferHeartbeat/browser
  - Additionally CMS has identified 3 groups in opposing time zones to monitor transfers
  - Use new EDM data sample
    - In April we only expect to have 5 days worth of unique new EDM data

In addition to low level transfers, CMS would like to demonstrate the bursting nature of Tier-1 to Tier-2 transfers

- Demonstrate Tier-2 centers at 50% of their available networking for hour long bursts

Accessing the Data



Job Submission

CMS calculates roughly 200k job submissions per day in 2008

- Calculation makes a lot of assumptions about the number of active users and kind of data access.

Aim for 50k jobs per day during 2006.

- CMS will begin transitioning to gLite 3.0
  - This is slow going. The evolution of the existing services generally work as expected, but the new services are somewhat unstable.

A larger number of application failures come from data publishing and data access problems than from problems with grid submission

- Need new event data model and data management infrastructure to have reasonable test
- About ready to deploy with PhEDEx
  - Publishing components are essentially eliminated

Currently CMS has only basic data access applications.

- Data skimming and selecting applications will be available by June

Job Submission Schedule

We expect a analysis grid job submitter capable of submitting to new event data model jobs

- ➡ Prototype exists: the configuration files are very different for CMSSW

By April we will enable user submissions to the limited sample of new EDM data

- ➡ At this point we can start functionality tests of gLite 3.0

During the first two weeks of June during the CMS focused section we would like to hit 25k jobs per day.

July and August we will be performing simulation at a high rate

Timeline (4/5)

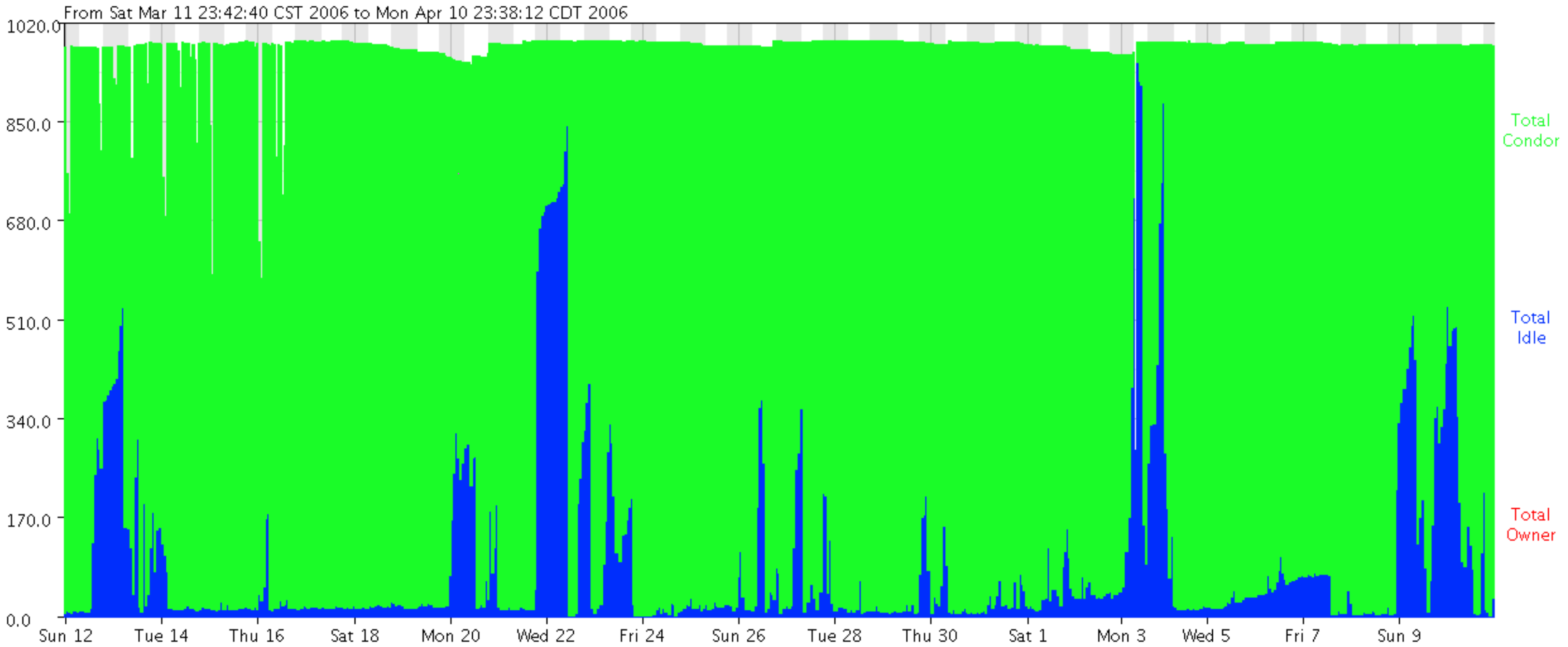
July August: CSA06 Preparation

- During July and August we expect to produce roughly 25M events per month across all of CMS
  - Roughly 2.5 times higher than we have currently demonstrated
  - 25% of the rate we expect in 2008
- We will also be exercising transfers and job submissions
  - Looking to improve efficiency
  - Deal with problems discovered during the SC4 concentrated activity
  - Improving grid efficiency. Continued cataloging and diagnosis of failure modes

Procurements

The CMS farm runs on average of 90% utilized

- ➡ Some of the missing periods have been failures of the negotiator due to an extreme number of requests waiting to be scheduled
  - In order to perform the physics validation, maintain production running, and keep user activities we need to start node procurement now
    - Goal is first rack of CPU in time to participate in production



Timeline (5/5)

September 15 - Oct 31 CSA06

- Computing and Software Analysis challenge is the last challenge before the start of data taking
  - Designed to exercise all aspects of the computing infrastructure at a rate of 25-50Hz (25% of 2008 trigger rate)
    - Reconstruct data at CERN
    - Distributed to Tier-1 centers
    - Exercise rereconstruction at Tier-1
    - Execute selection and skimming applications
    - Replicate data streams and selected datasets to Tier-2 centers
    - Produce simulation at Tier-2 and archive to Tier-1s
    - Execute prototypical alignment and calibration applications and feedback
    - Provide access to distributed conditions data
  - While we anticipate a lot of people participation in the challenge for analysis activities, we are trying not to overload this activity with the PTDR V3 work
    - Demonstrate start-up scenarios with CMSSW

Procurements

CMS milestone calls for 50% system complexity by the end of FY06

- ➡ Final year of the procurement doubles to 100% complex by the end of 07

Our contribution to the CSA06 is 300 modern processor nodes

- ➡ The computing model calls for 150TB of disk for every 1MSI2k
  - To maintain these ratios we need roughly 400TB of disk commissioned at FNAL to complete the challenge

In order to provide resources for the challenge and the PTDR we need to have the new nodes commissioned by the end of August

- ➡ We also need not to overly stress the computing facility to meet the 2006 milestone during CSA06
- ➡ Makes sense to start the node procurement now
  - First rack in first two weeks of June
  - Remaining racks by beginning of August